Carnegie Mellon

Ambient Intelligence Lab

Spatiotemporal Bayesian Prediction Model

Yang Cai, Ph.D.
Ambient Intelligence Lab
Carnegie Mellon University
ycai@cmu.edu

Collaborators

Karl Fu, Carnegie Mellon
Xavier Boutonnier, Carnegie Mellon
Daniel Chung, Carnegie Mellon
Mohamed Abid, Carnegie Mellon
Mel Siegel, Carnegie Mellon

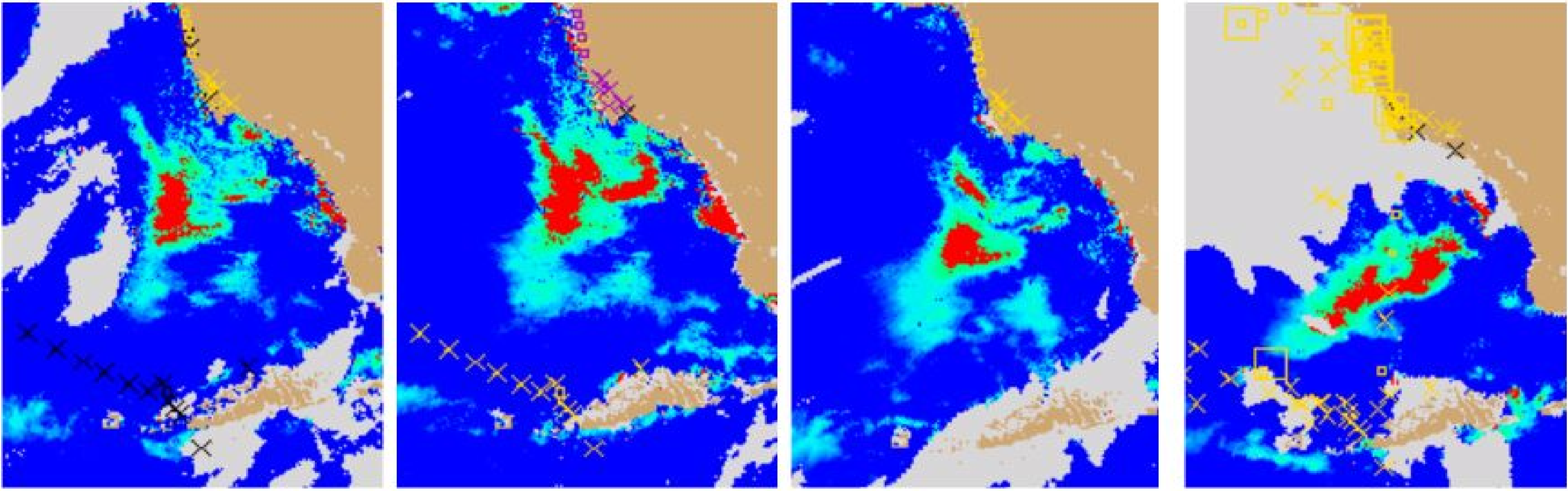
Richard Stumpf, NOAA (Co-I)
Timothy Wynne, NOAA
Mitchell Tomlison, NOAA

James Acker, GSFC
Yonxiang Hu, LaRC

Cynthia Heil, FWRI
Andy Moore, Google
Eric Mararet, Act Corp.

“Red Tide”

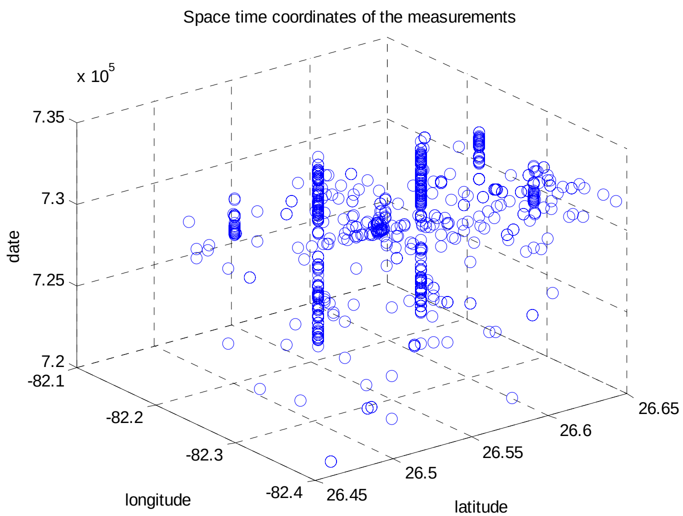
A spatiotemporal problem



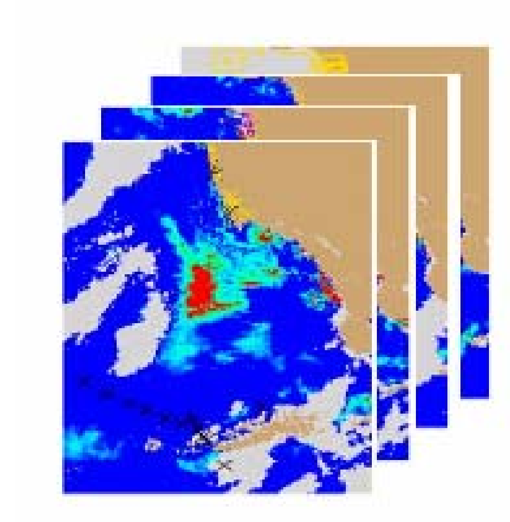
Images above show a harmful algae bloom (HAB), highlighted as chlorophyll anomaly, drifting along the southwest Florida coast in December 2001.


PYREX
Florida Red Tide
Karenia brevis


T3
T7

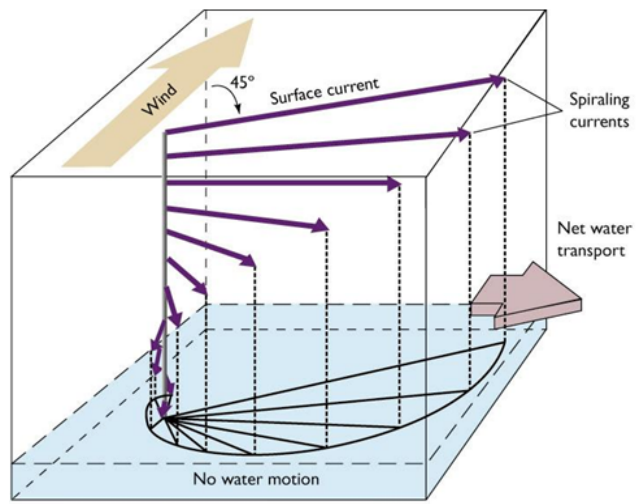


Model



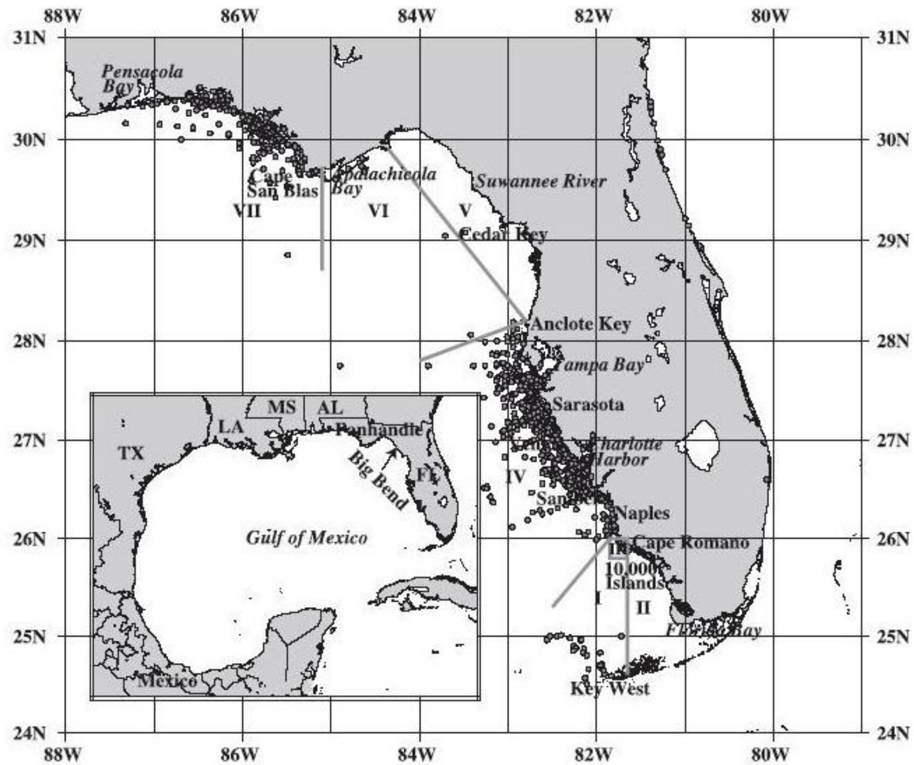
In-Situ sensor data
(cell count)

Satellite images
(SeaWiFS)



Physical observations (e.g.
current, salinity, temperature)

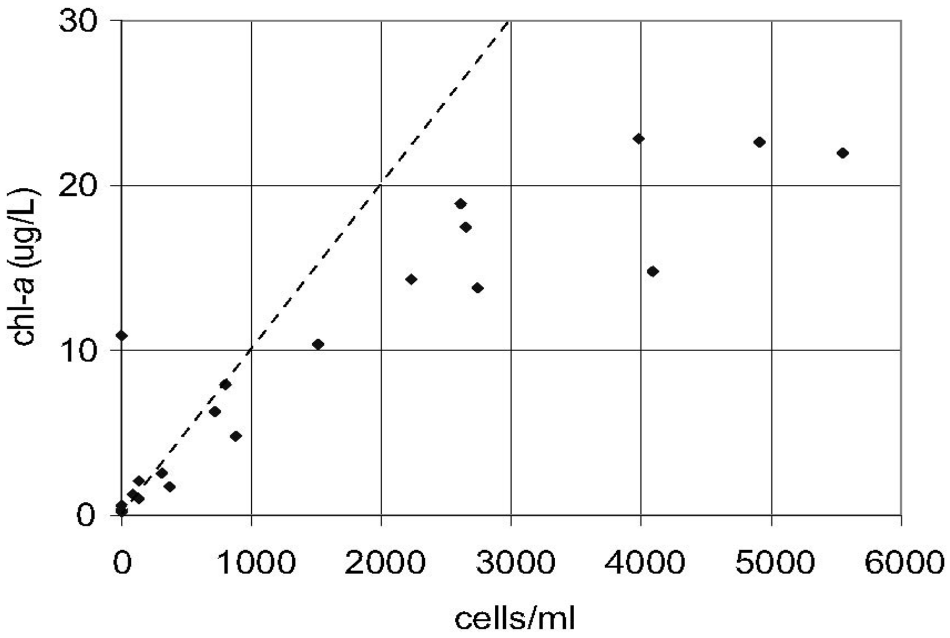
Cell Count Data



West Florida regions divided by NOAA scientists

Correlation Study

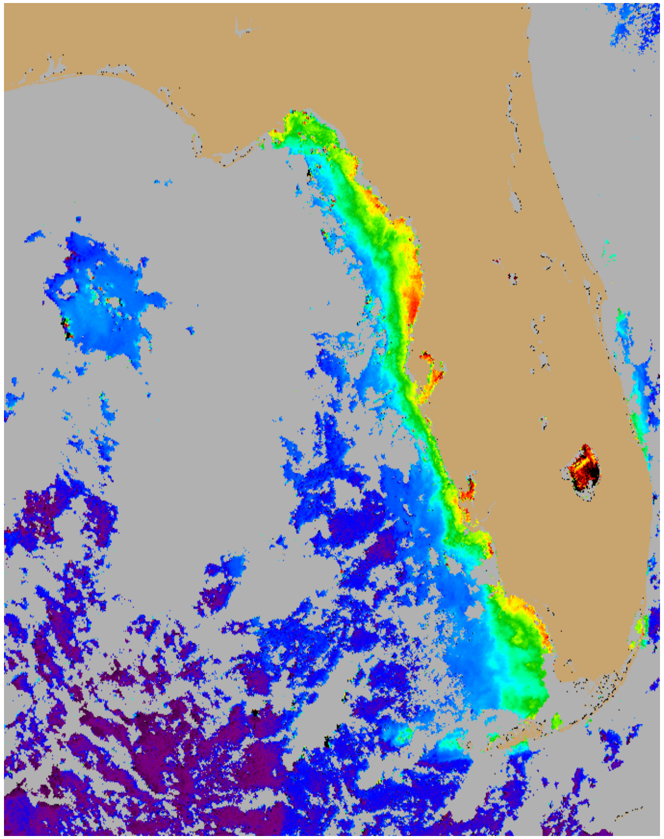
Use chlorophyll as a surrogate for Karenia Brevis blooms (NOAA)



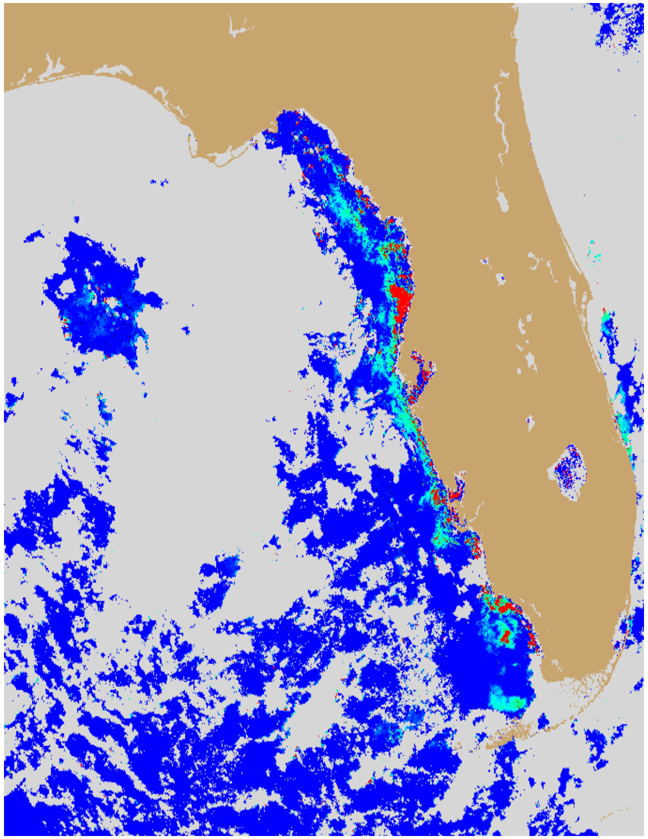
References:

Tomlinson, M.C., R.P. Stumpf, V. Ransibrahmanakul, E.W. Truby, G.J. Kirkpatrick, B.A. Pederson, G.A. Vargo, C. A. Heil., 2004. Evaluation of the use of SeaWiFS imagery for detecting Karenia brevis harmful algal blooms in the eastern Gulf of Mexico. Remote Sensing of Environment, v. 91, pp. 293-303.

SeaWiSF Database

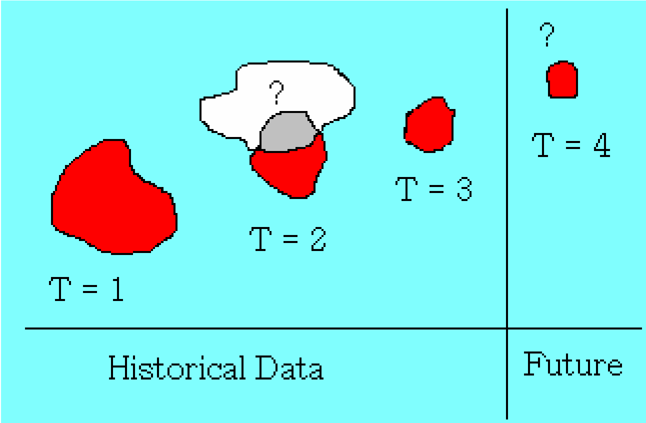


Chlorophyll channel



Anomaly channel

Scientific Questions



Given databases of historical data and current physical and biochemical conditions, how to predict the occurrence of the target at a particular time and location?

Vision + Mining

Vision:

- Spatial Density Filter
- Correlation Filter and Particle Filter
- Mutual Information

Mining:

- Spatiotemporal Neural Network
- Spatiotemporal Bayesian Model
- Periodicity Transform

Raw image without mapping

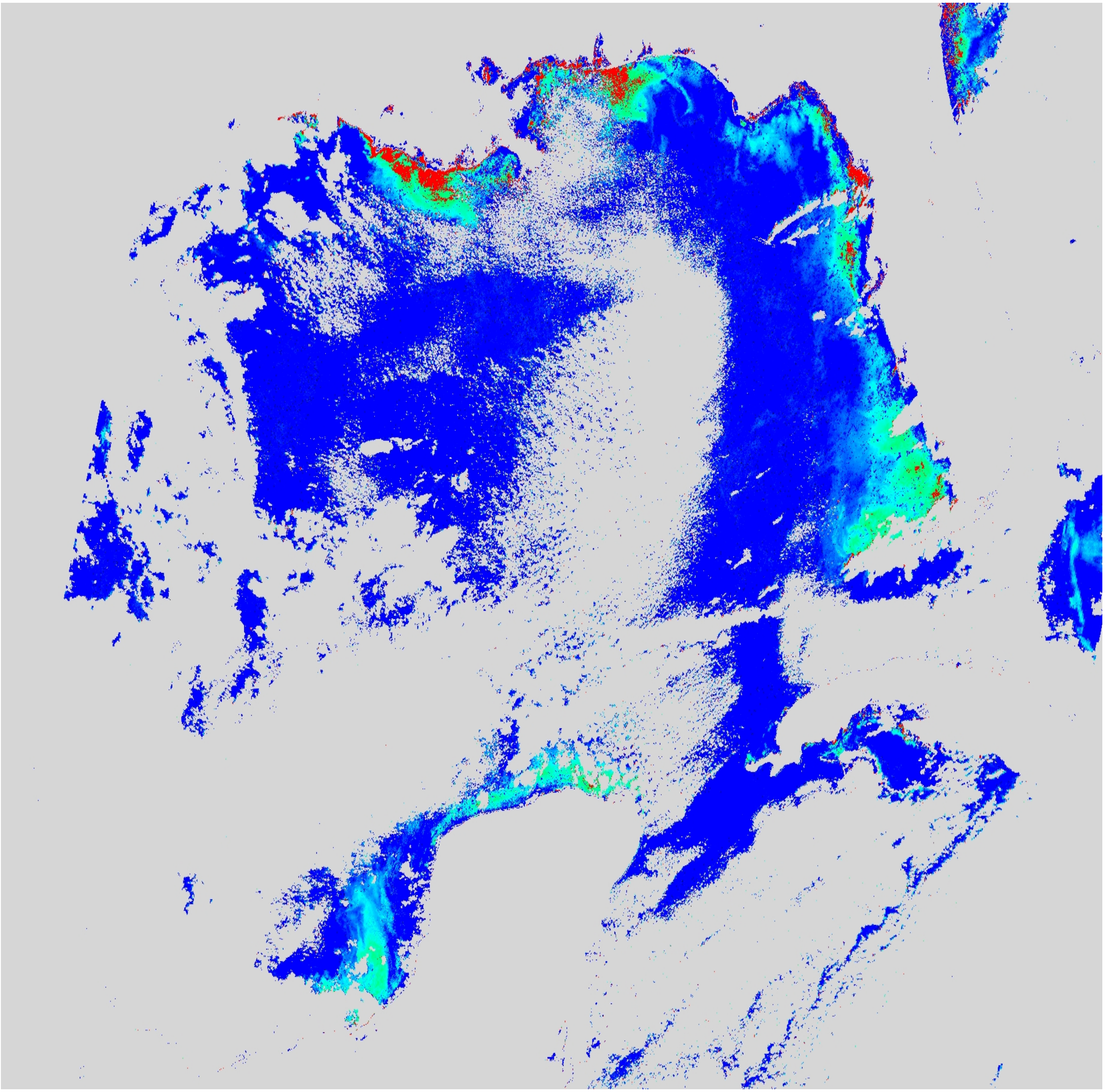
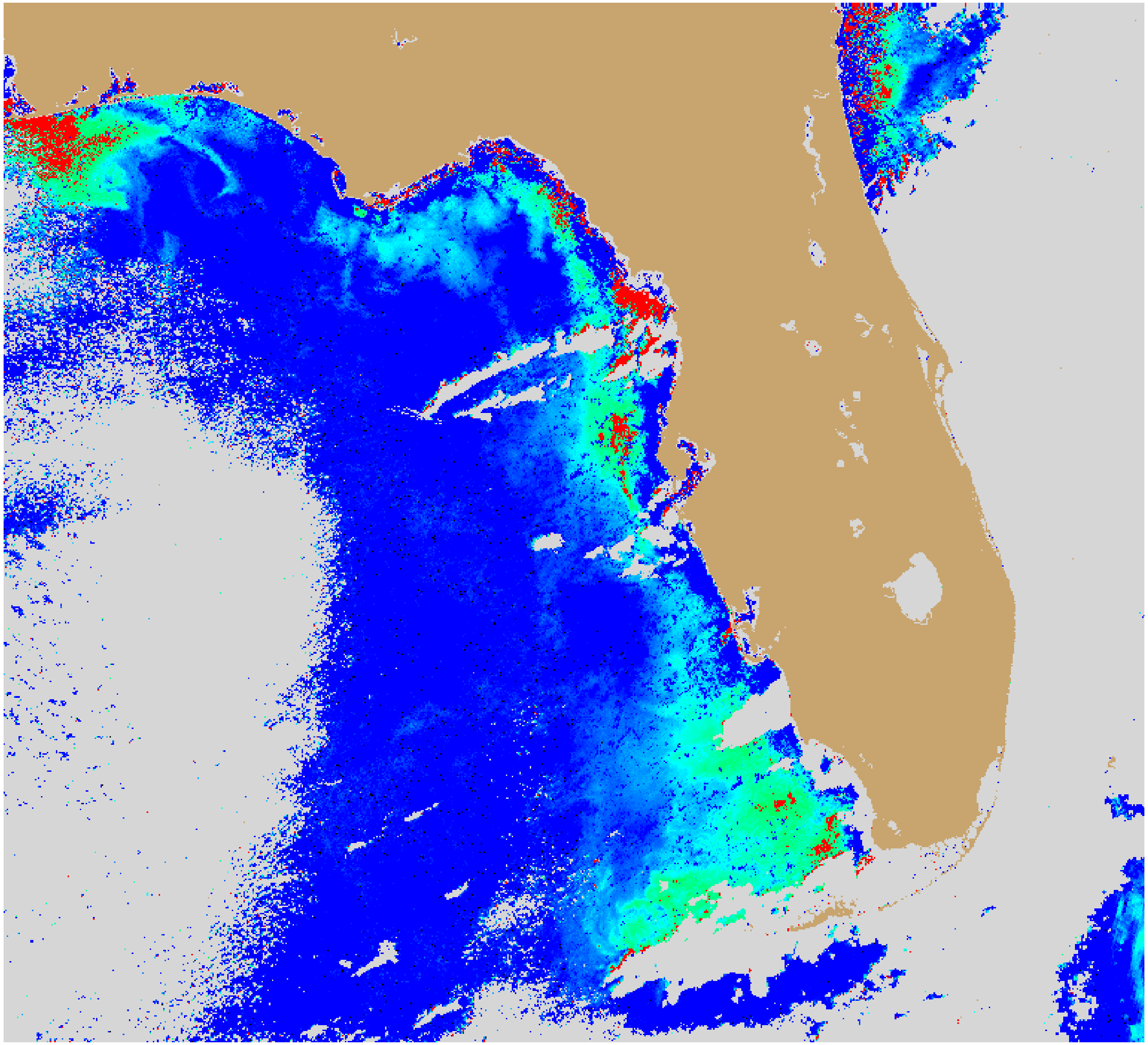
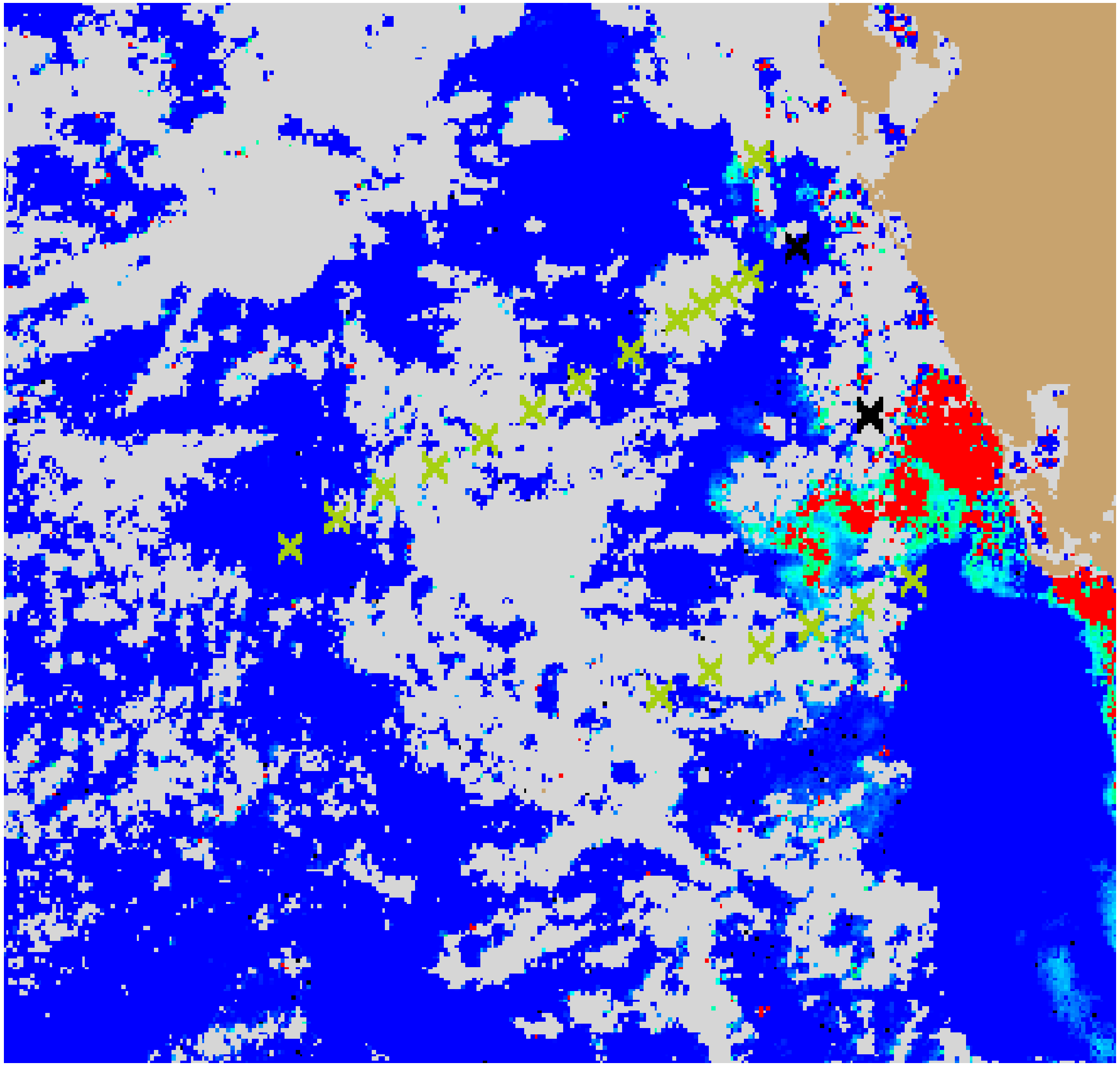


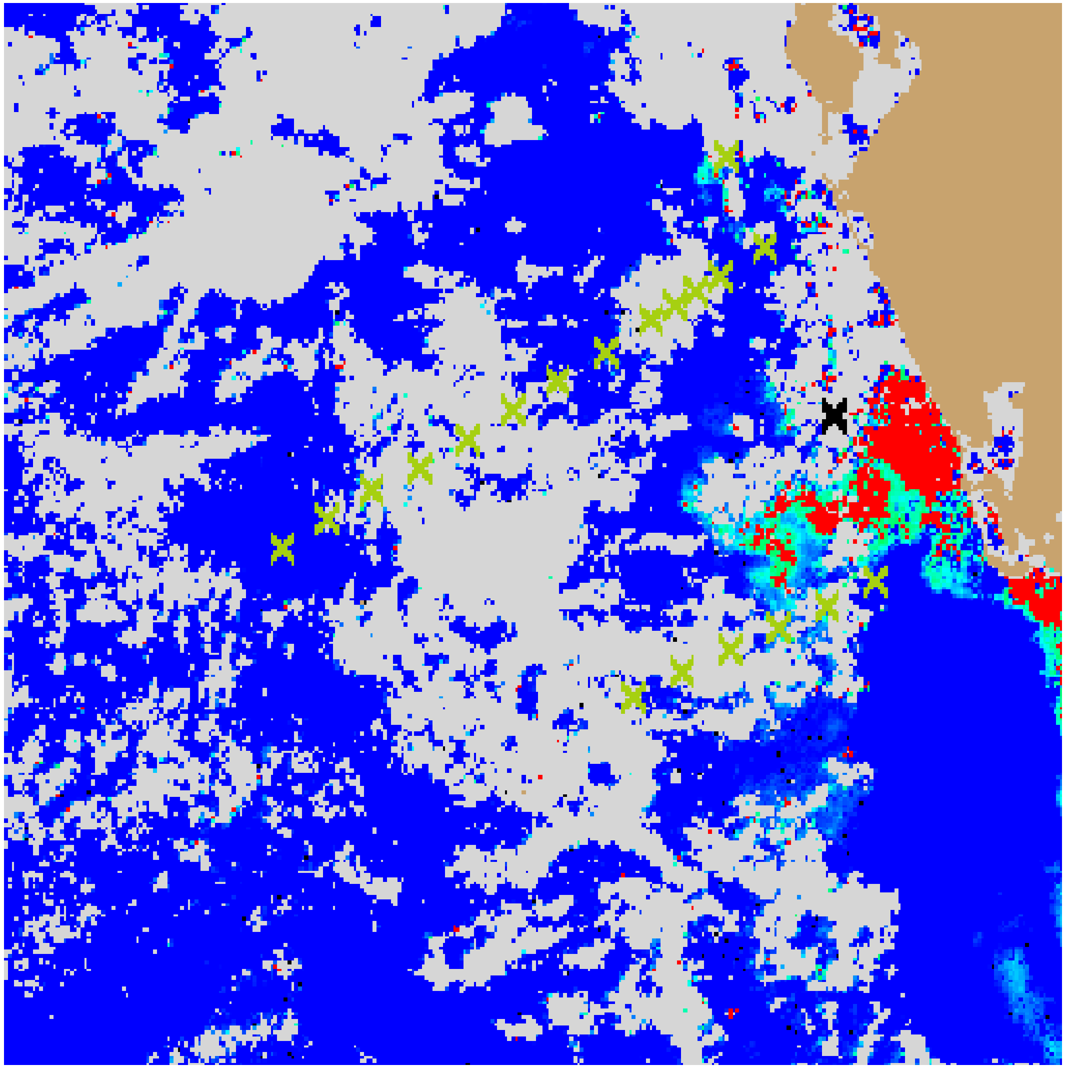
Image after cropping and remapping



Ground Truth for October 5, 2000



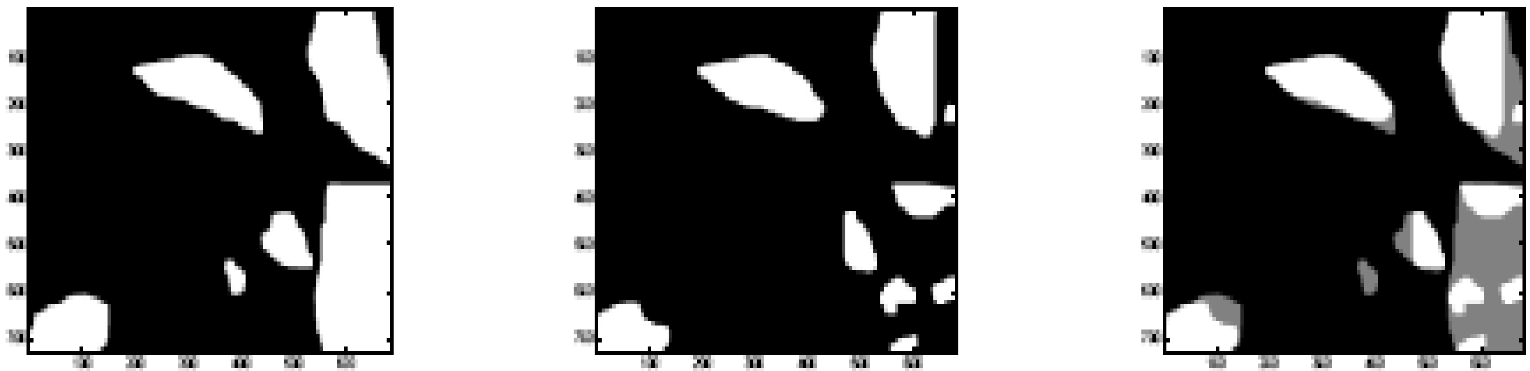
Predicted for October 5, 2000



Missing data recovery

Over 85% of images contain clouds. So we have significant amount of missing data in visible satellite images.

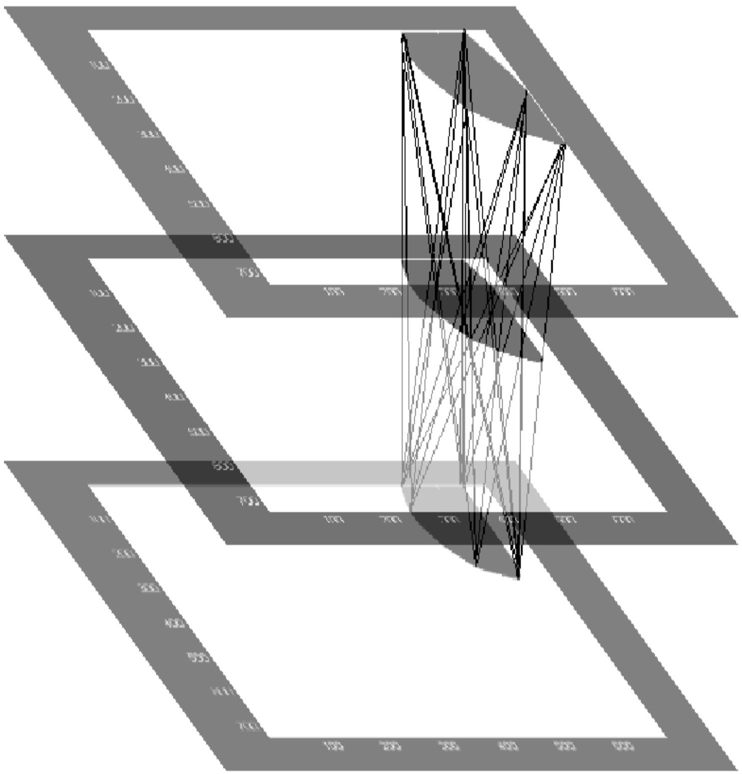
Missing data recovery with
shape interpretion



1. Concavity of objects?

2. Which is which?

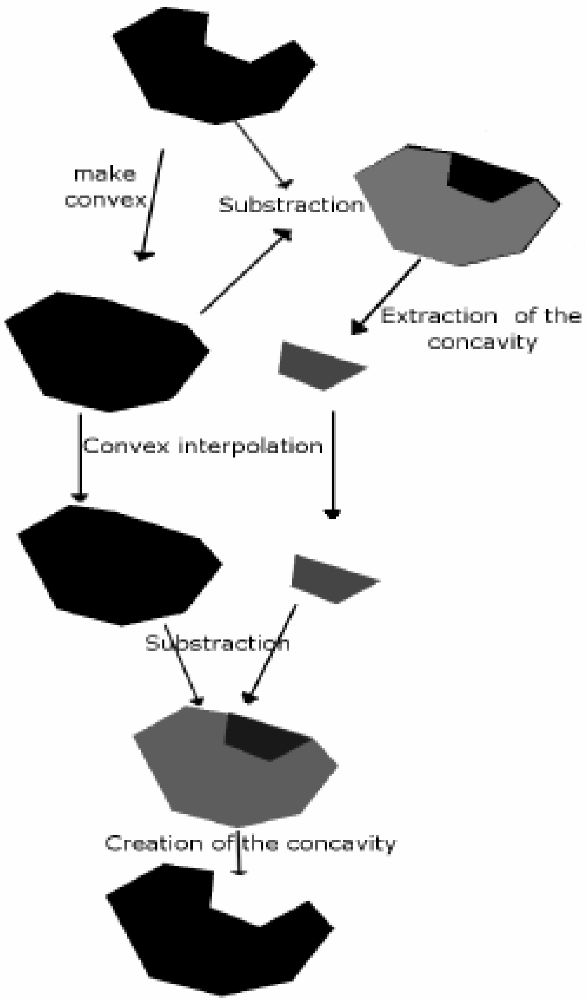
Interpolation of a convex object



We take all the points of the contours of the marginal objects and by linear interpolation calculate the position of the interpolated point.

The Hull Convex of the interpolated points gives the contour of the interpolated convex object.

Work around concavity



1. First we extract the concavity.
2. Then we interpolate the object and the concavities.
3. Then we remove the part corresponding to the interpolated concavity from the interpolated object

Results

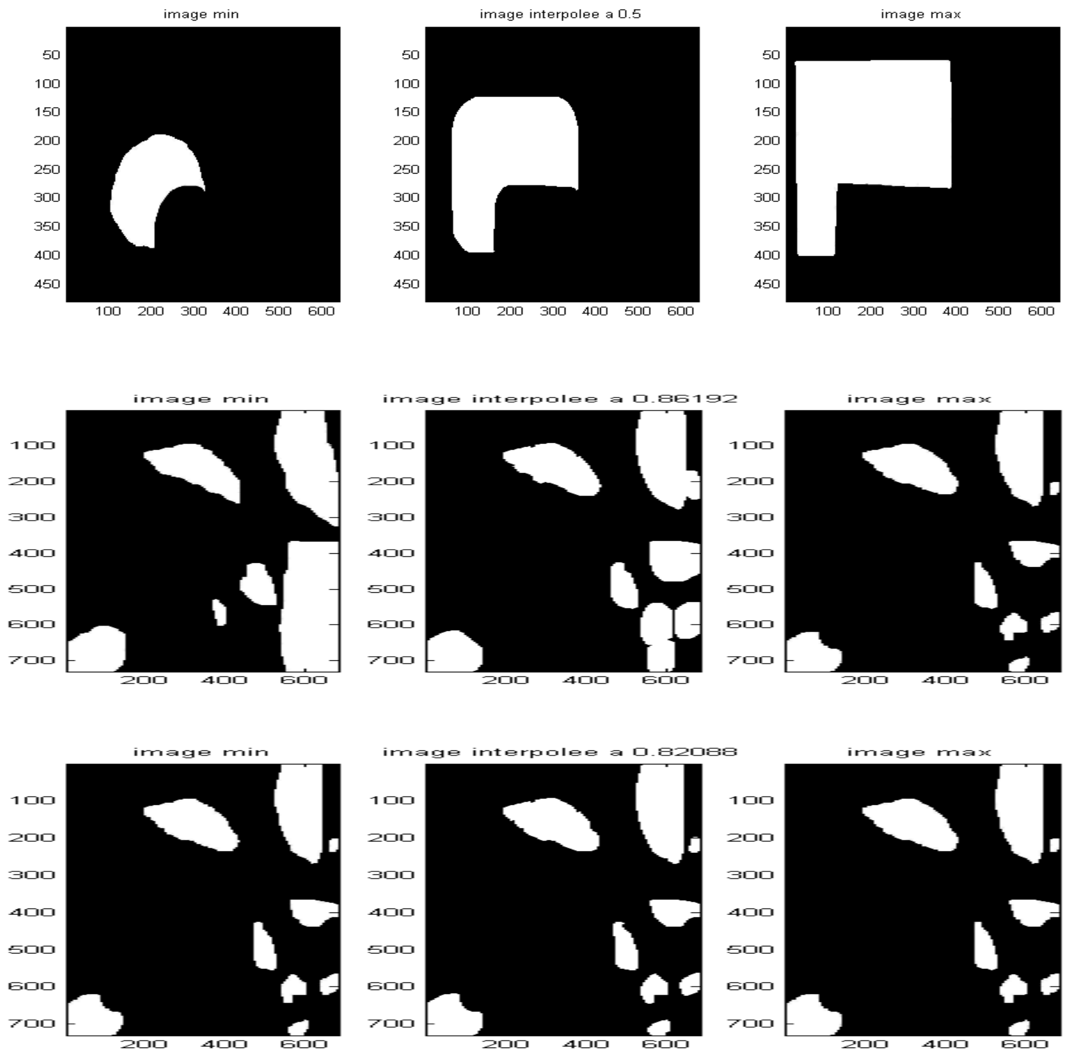


image min
image interpolee a 0.5
image max
image min
image interpolee a 0.86192
image max
image min
image interpolee a 0.82088
image max

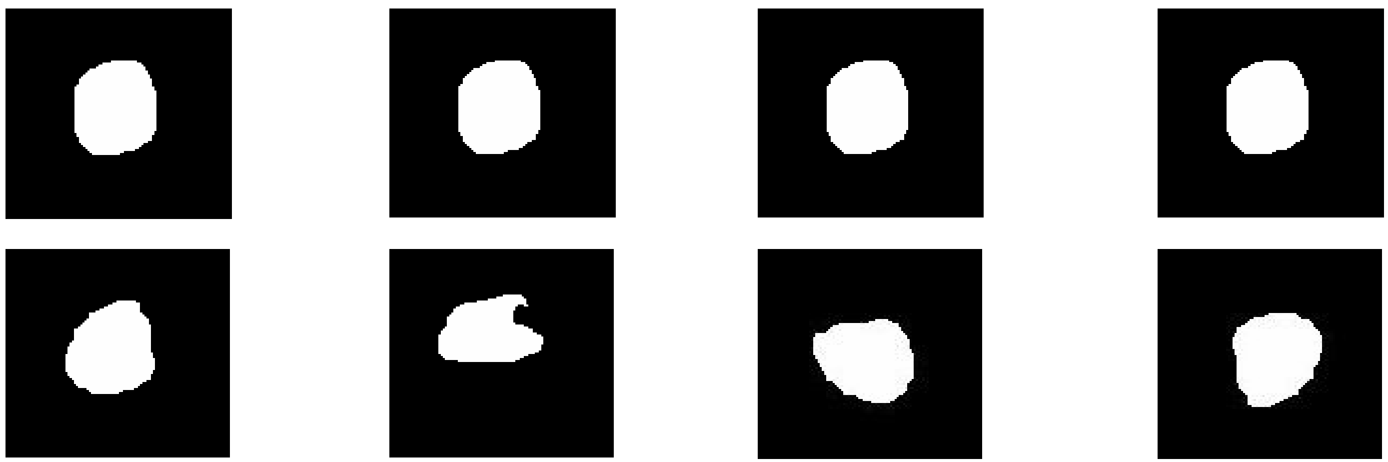
Missing Data Recovery with Mutual Information

Mutual Information measures dependency between two variables. If *X* and *Y* are independent, then *X* contains no information about *Y* and vice versa, so their mutual information is zero.

$$I(X;Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) \log \frac{p(x,y)}{f(x)g(y)}$$

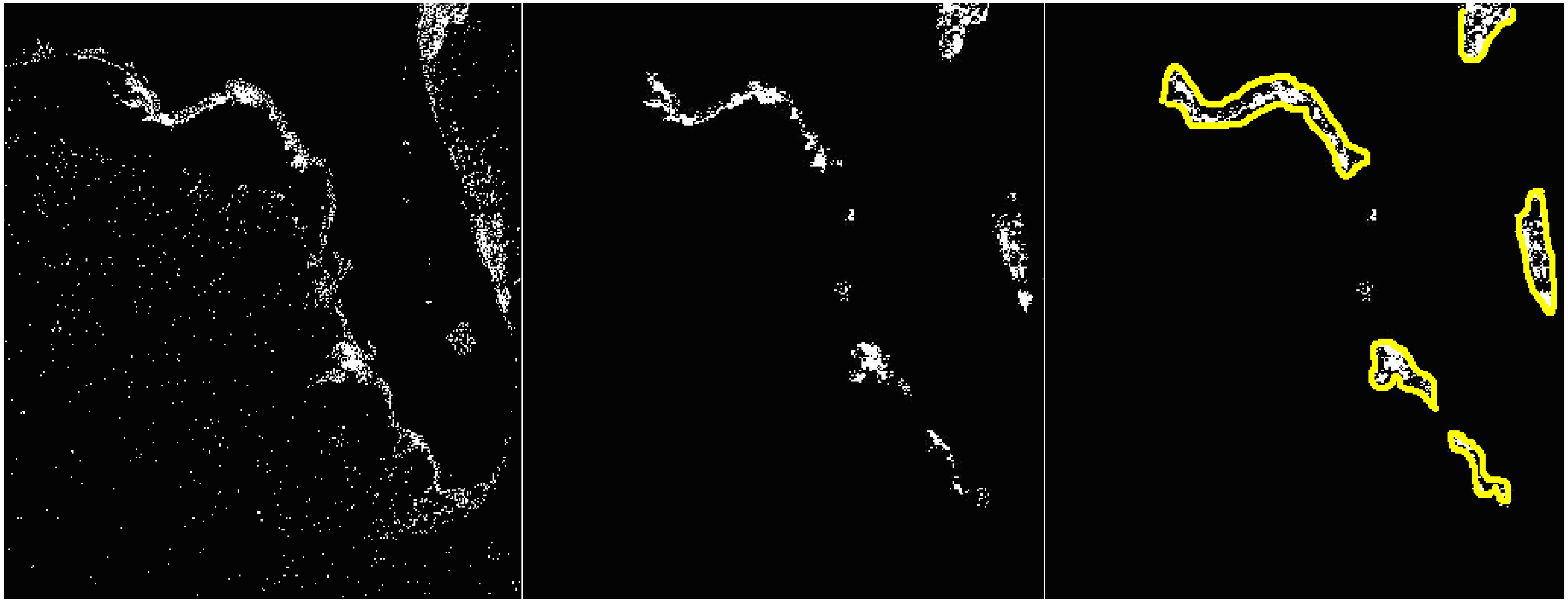
where *p* is the joint probability distribution function of X and Y, and *f* and *g* are the marginal probability distribution functions of X and Y respectively.

Examples of mutual information measurement

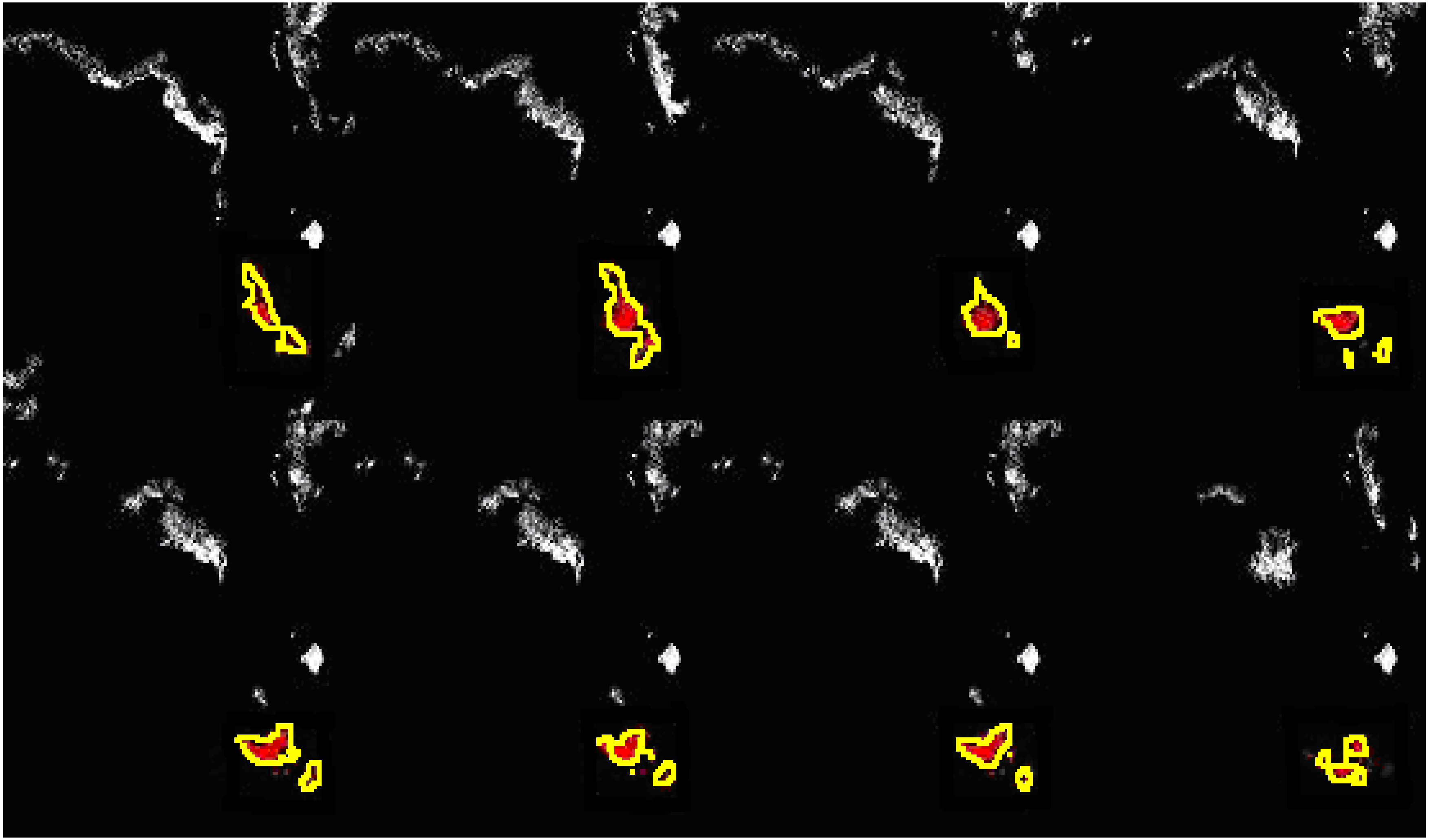


MI = 0.5122   MI = 0.4645   MI = 0.4887   MI = 0.4869

Shape Grouping Results

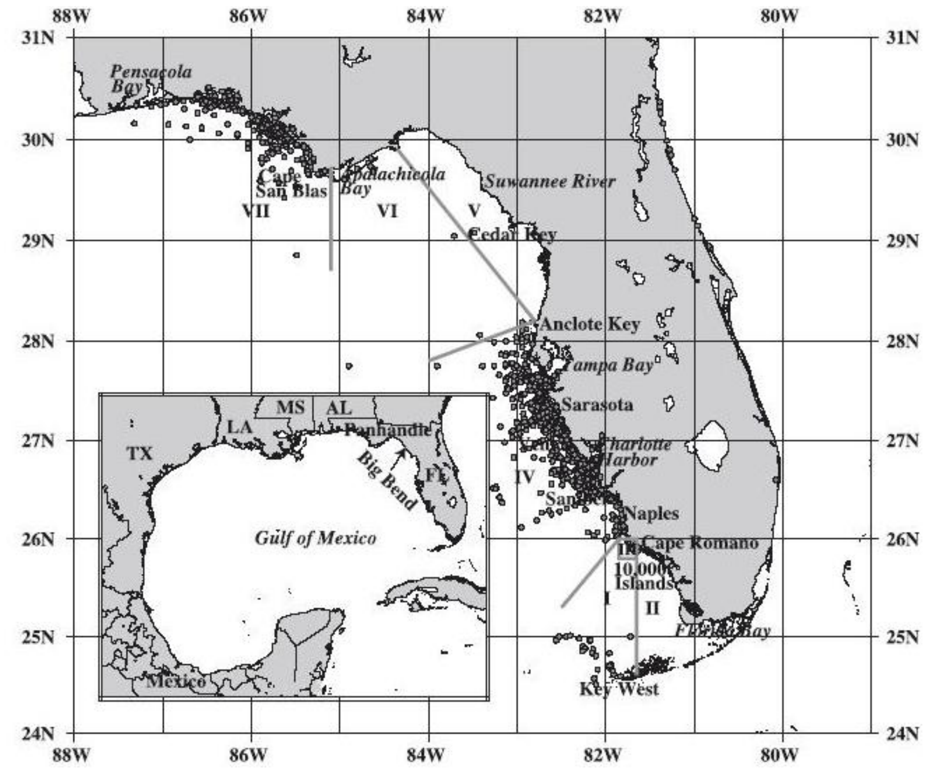


Recovered images



Recovered sample frames

Object tracking across multiple zones



West florida regions divided by NOAA scientists

Trans-region HAB movements

From the period of September 28th 2000 to December 2nd 2002

Date	Path ID	Transition (Region to Region)
10/3/2000	2	7 to 6
12/22/2001	35	4 to 1
1/27/2002	51	4 to 5
7/27/2002	60	2 to 4
8/13/2002	69	4 to 2
8/15/2002	69	2 to 1
8/17/2002	69	1 to 2
10/27/2002	82	7 to 6

Bèzier Curve for Trajectory

Bèzier Curve is a way that computer stores a curve in its memory. It consists of two end points and zero or more control points in between. Each point on the curve can be determined by B(t).

$$B(t) = \sum_{i=0}^{n} C_i^n P_i b_{i,n}(t) \text{ , t} \in [0, 1]$$

where the polynomials

$$b_{i,n}(t) = C_i^n (1-t)^{n-i} t^i$$

P_i is called the control points which will be the centers of gravity in our case.

Trajectory of a HAB movement

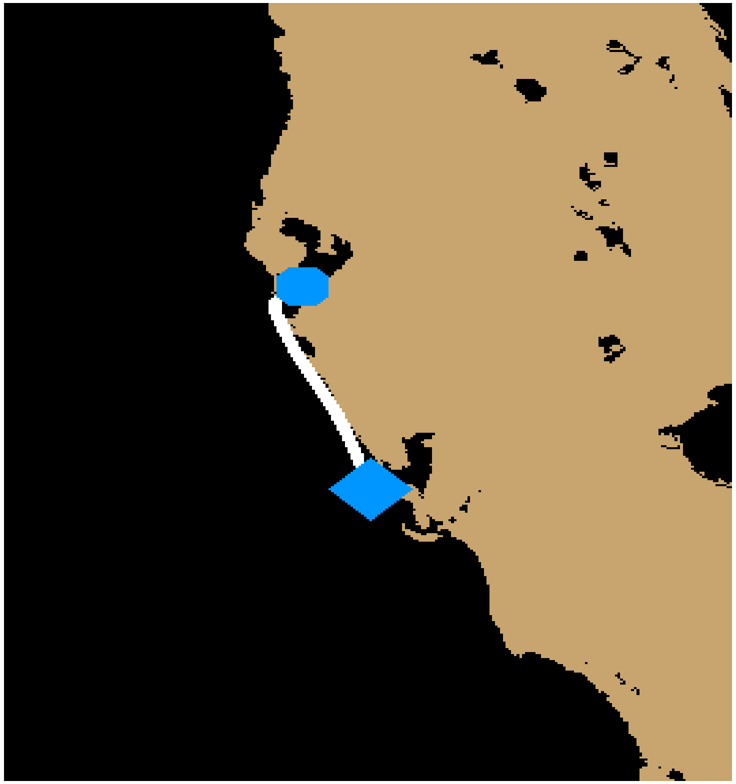


Figure 2 Trajectory of a HAB
Circle (start): June 15st 2001
Square (end): July 29th 2001

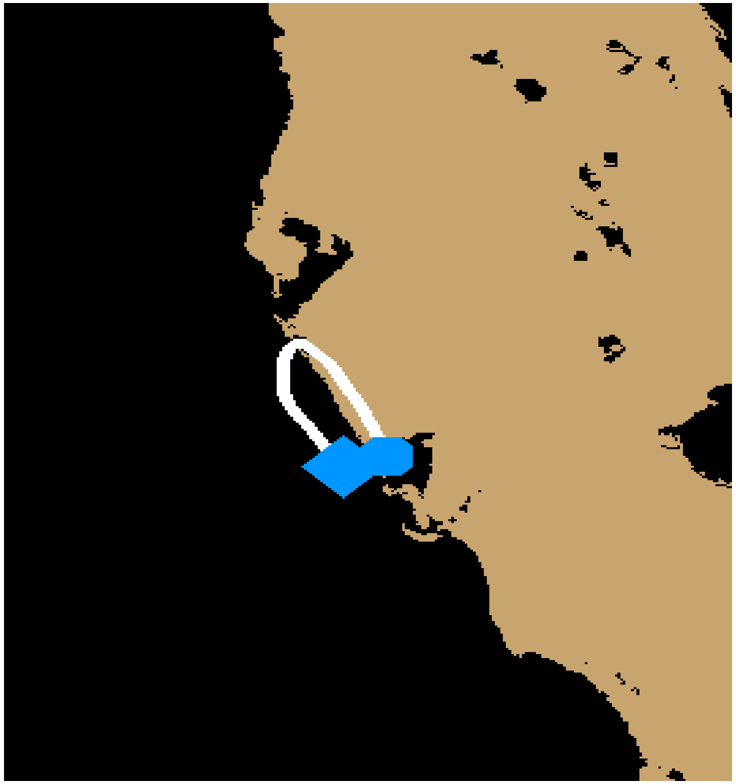
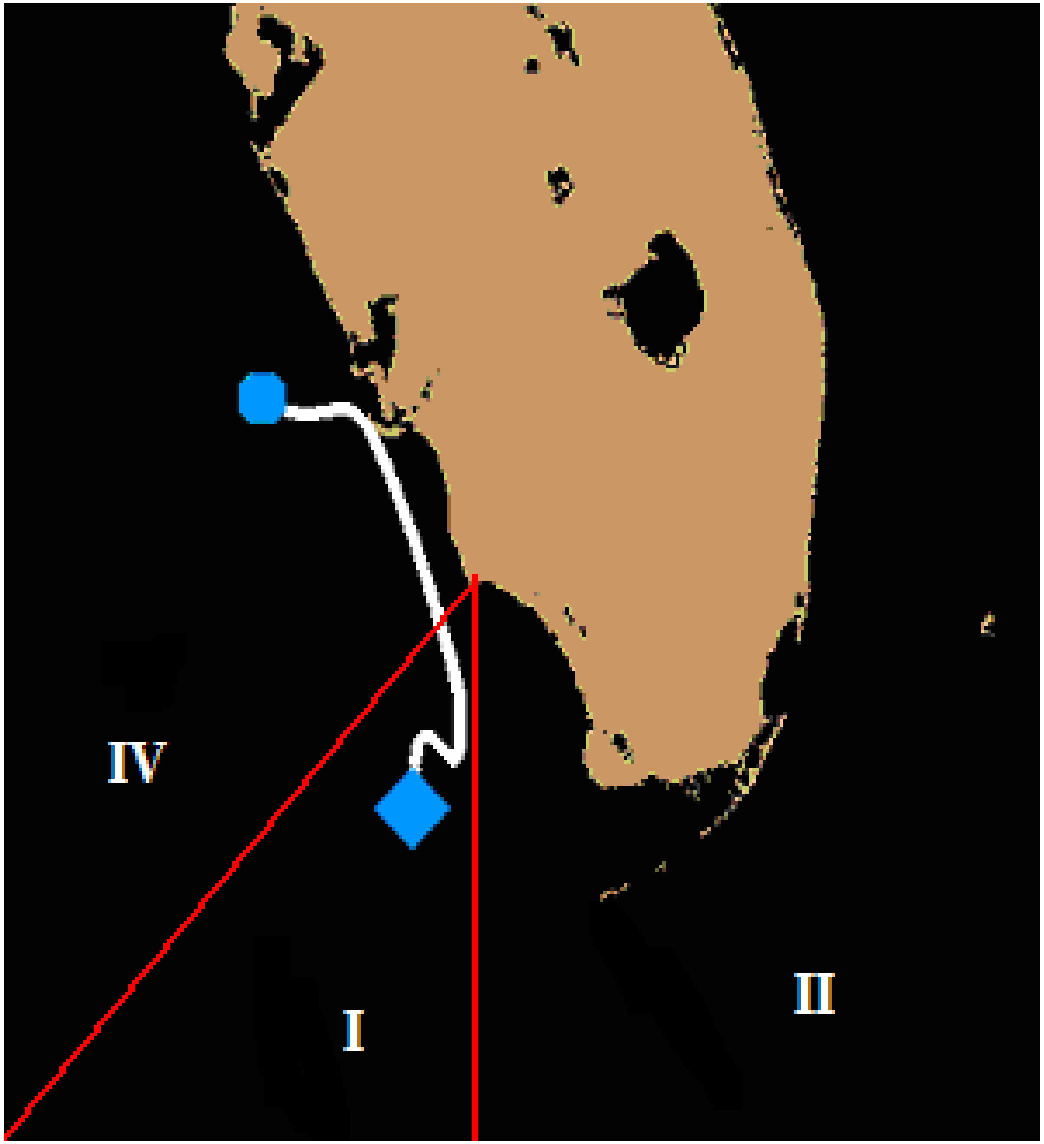
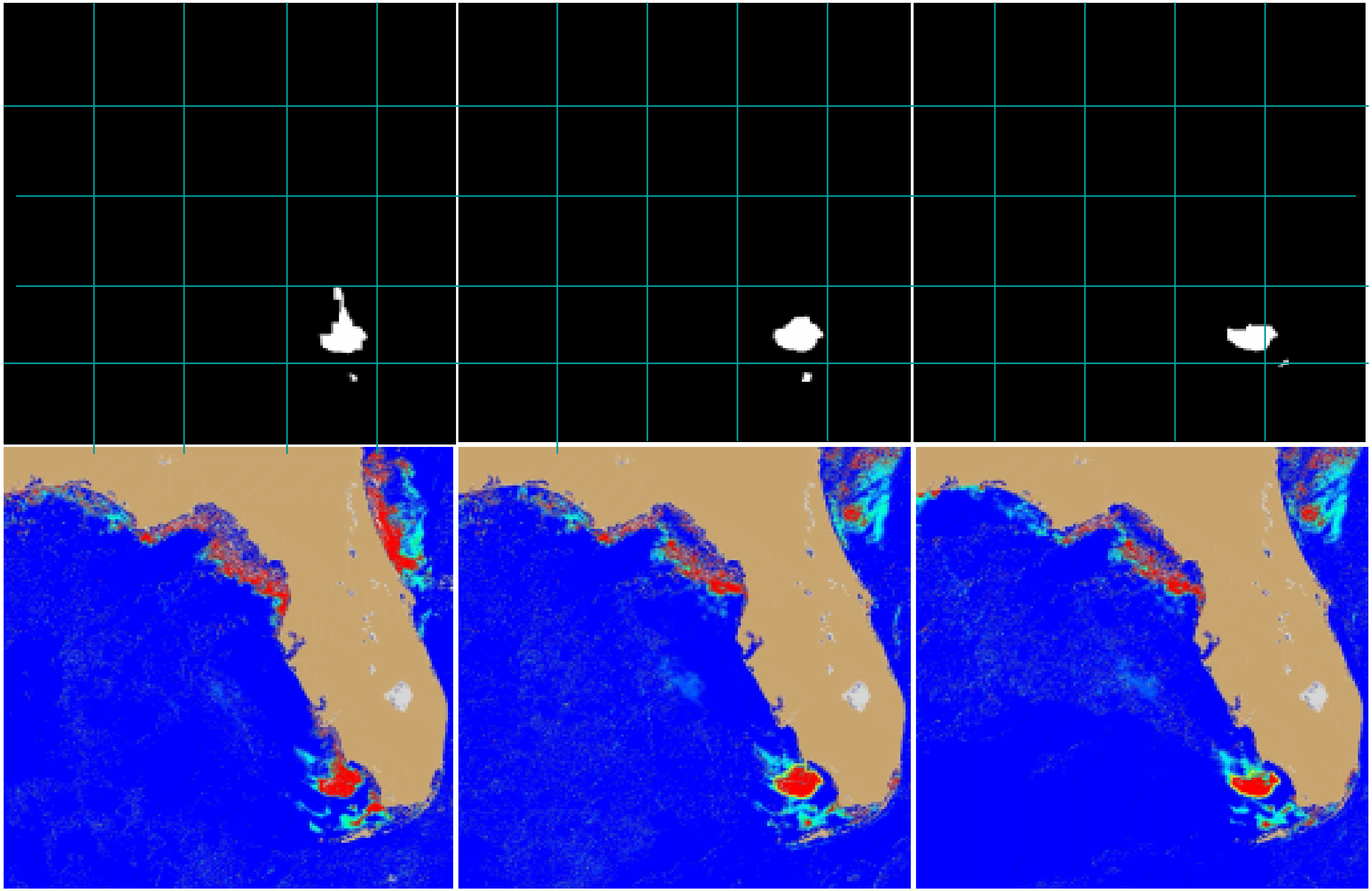


Figure 3: Trajectory of a HAB
Circle (start): August 2th 2001
Square (end): October 20th 2001

Compuetrized trajectory tracking

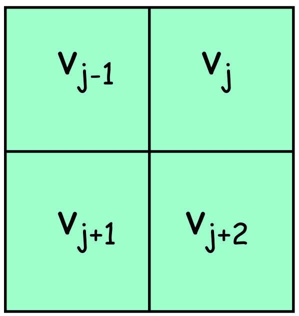


Marked surface object in a cellular automata grid



Spatiotemporal Bayesian prediction model

Assume that all evidences (e_1,...e_i) are independent.
The model is to find the maximized probability for a state to be true.



$$v_j = \arg\max_{v_j \in V} P(v_j) \prod_{k=1}^{i} P(e_k \mid v_j)$$

v_j = state of lot j

$P(v_j)$ = prior probability of state v_j

$P(e_k | v_j)$ = likelihood of the evidence e_k being true

i = total number of evidences

V = set of states

Naïve Bayesian Calculation

$$P(v_j) = \frac{N_v}{N}$$

N_v ← total number of training instances with state v_j

N ← number of training sets.

$$P(e_k \mid v_j) = \frac{P(e_k \cap v_j)}{P(v_j)} = \frac{N_c}{N_v}$$

N_c ← number of training instances with evidence e_k and state v_j

Handling sparse data

Sparse data yield inaccurate results, i.e. N_c is small

Better version: $$P(e_k \mid v_j) = \frac{N_c + mp}{N_v + m}$$

m is the constant to enlarge the sample size

p is the prior estimate of the probability such that $p = \frac{1}{r}$

where r is number of values that e_k can take.

(assuming uniform prior)

Convert field data into the model

Given historical data about the time, location, and the presence of HAB for each entry, we assume a location (x_0, y_0) and a time t_0, equation (1) is used to predict whether HAB is present or not. $c \in \{0,1\}$, '0' represents cell count being less than 5,000, and '1' represents cell count being greater than or equal to 5,000.

$$SB(x_0, y_0, t_0) = \arg\max_{c=0,1} P(c) R(x_0|c) R(y_0|c) R(t_0|c) \quad (3)$$

$R(a_0|b)$ Denotes $P(a|b)$ where $a = \{s \mid a_0 - \mu \leq s \leq a_0 + \mu\}$ and μ is a positive constant. Using this ranged probability, we can spatially group our training set.

Use Images as Evidences

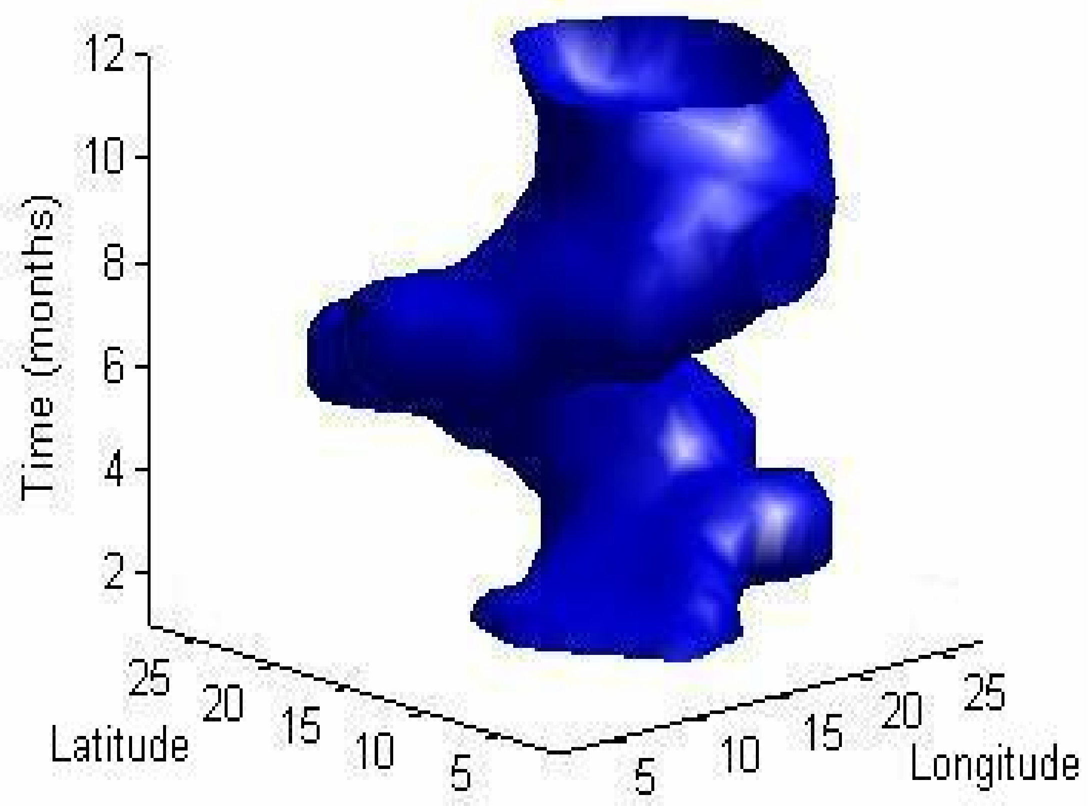
The grid alignment with other data.

Recursive Bayesian model: when a new image is added, only one additional term is multiplied onto the equation.

$$SB(x_0, y_0, t_0, \mathrm{I}) = \arg \max_{c=0,1} P(c) R(x_0|c) R(y_0|c) R(t_0|c) \mathrm{P}(\mathrm{I}_{x,y}|\mathrm{c}) \quad (5)$$

With image data alone, the prediction accuracy increases about 5%.

Visualization of the prediction model



Results with 2,384 samples vs. 188 samples

ground truth = cell counts

Table 1. Our prediction methods

Method	False positive	Confirmed positive	False negative	Confirmed negative	Sum	Positive detection	Positive Accuracy	Negative Accuracy
Image only	44	17	6	306	373	86.60%	73.91%	87.43%
SB[1] w/o SDT[2]	161	423	142	1658	2384	87.29%	74.87%	91.15%
SB w/ SDT	176	445	120	1643	2384	87.58%	78.76%	90.32%
SB w/ SDT & Int[3]	166	441	124	1653	2384	87.84%	78.05%	90.87%
SB w/ SDT & Or[4]	173	445	120	1646	2384	87.71%	78.76%	90.49%

1 Spatiotemporal Bayesian Model
2 Sparse Data Treatment
3 Using interpolated images
4 Using original images

Table 2. The tabulated prediction reference results from a published paper

Method	False positive	Confirmed positive	False negative	Confirmed negative	Sum	Positive detection	Positive Accuracy	Negative Accuracy
Reference Results	5	36	23	124	188	85.10%	61.02%	96.12%

Positive accuracy is the percent of the cases in which HAB is present and the model predicted correctly.
Positive accuracy = confirmed positive / (confirmed positive + false negative)

Positive detection is the percent of ALL predictions that are correct.
Positive detection = (confirmed positive + confirmed negative) / (sum)

Fusion of Multiple Databases

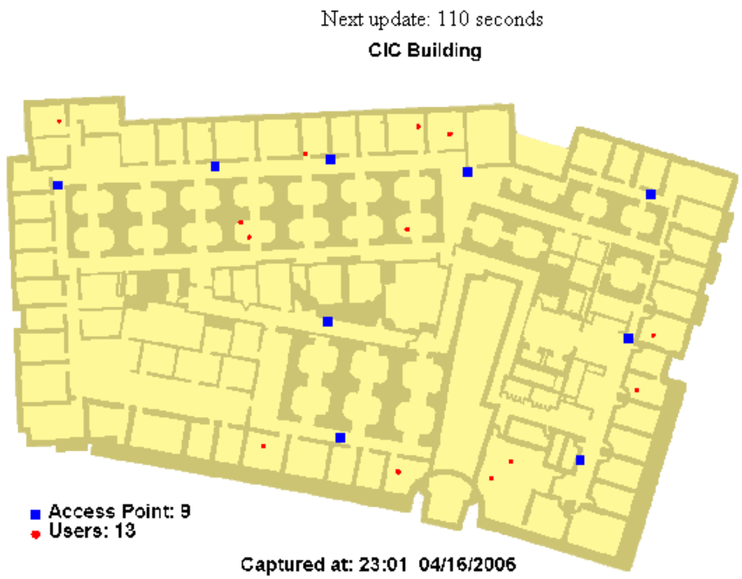
- SeaWiFS (8 years)
- Cell count (50 years)
- Wind
- Temperature
- Salinity

On-going technical infusion

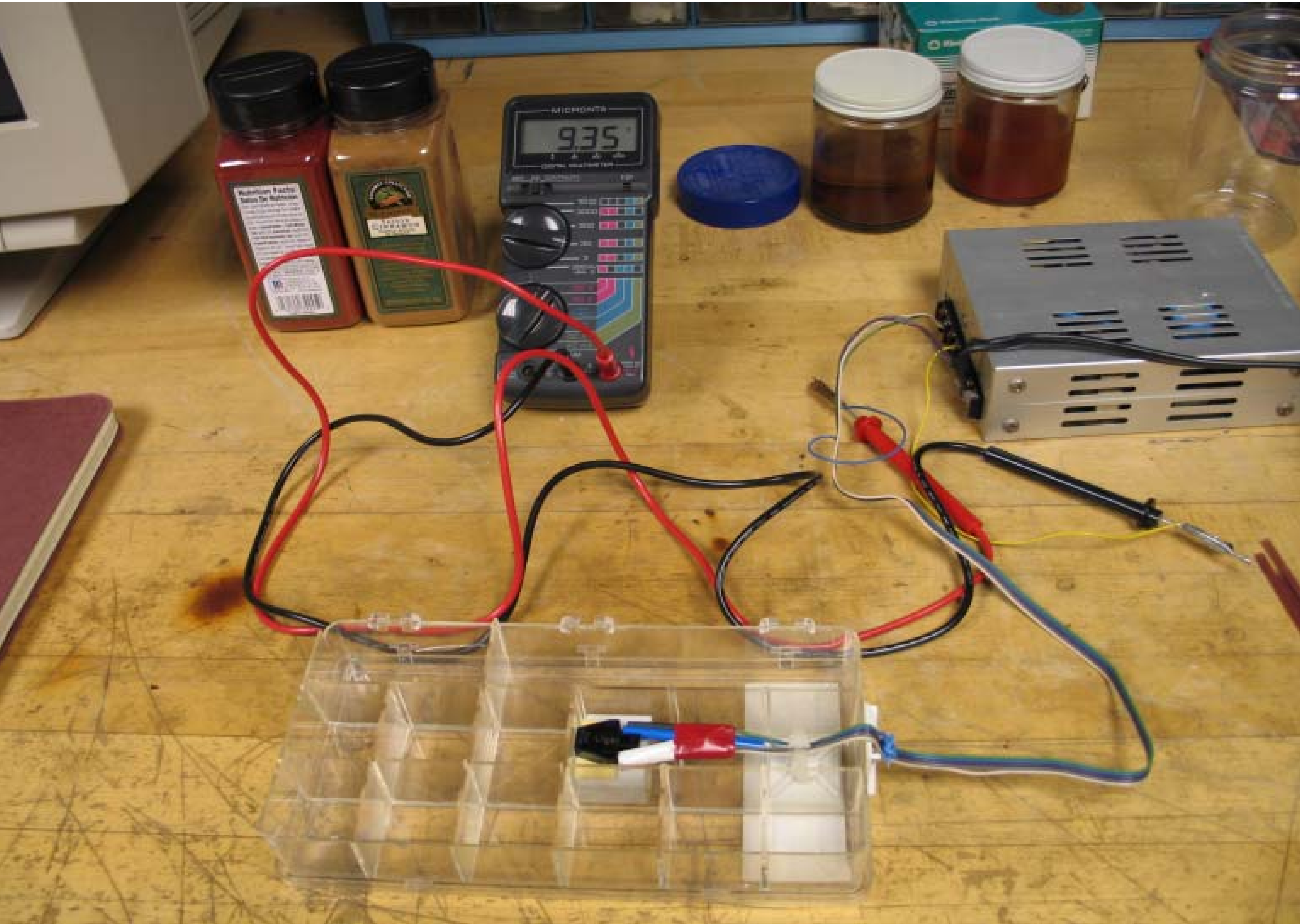
- NOAA -> research toolkit
- Florida RWRI -> HAB BBS
- TIWG & Act Corp. -> remote data mining
- GSFC -> Gevanni system
- JPL -> fishery modeling
- LaRC -> sensor web

Projects for next generation data mining

- Multiphysics (Cellular Automata)
- Data-Mining-on-Chip (Neural Network on FPGA)
- Sensor Webs (CMUSky)
- Visualization of sensor web



HAB sensor prototype



Conclusions

1. The spatiotemporal Bayesian prediction model shows promising in positive detection of Harmful Algal Blooms based on the in-situ sensor data and satellite images. The fusion of databases increases the prediction accuracy (e.g. reduced the false alarms)
2. Vision algorithms are effective for recovering the missing data and tracking the surface objects. However, it’s challenging to register the noisy satellite images with the in-situ sensor data because of the different resolutions.
3. “Vision+Mining” technology enables automated process to verify hypotheses and real-time detection of objects. It would liberate scientists from manual analysis to computer-human interaction process. In this project, we tested 2384 cases on a PC versus 188 cases by hand.

Publications

1. **Y. Cai, R. Stumpf, etc. Spatiotemporal Data Mining for Prediction of Harmful Algal Blooms, International Harmful Algae Conference, Copenhagen, September 8-12, 2006**
2. Y. Cai, Y. Hu, Onboard Inverse Physics from Sensor Web, Proceedings of Space Missions and Challenges, SMC-IT, JPL, 2006
3. Y. Cai and K. Fu, Spatiotemporal Data Mining with Cellular Automata, Proceedings of International Conference of Computational Science, ICCS 2006, May 30, UK
4. Y. Cai, D. Chung, K. Fu, R. Stumpf, T. Wynne, M. Tomlinson, Spatiotemporal Data Mining with Micro Visual Interaction, submitted to Journal of Knowledge and Information Systems
5. Y. Cai, K. Fu, R. Stumpf, T. Wynne, M. Tomlinson, Spatiotemporal Data Mining for Monitoring Ocean Objects, submitted to NASA Data Mining Workshop, JPL, 2006
6. Y. Cai, Y. Hu, Sensory Stream Data Mining on Chip, submitted to NASA Data Mining Workshop, JPL, 2006
7. Y. Cai, (editor), Special Issue of Visual data Mining, Journal of Information Visualization, to be published by Elsevier, 2006
8. Y. Cai and J. Abascal, (editors), Ambient Intelligence in Everyday Life, Lecture Notes in Artificial Intelligence, LNAI 3864, to be published by Springer, April, 2006

Acknowledgement

The study is supported by NASA ESTO grant AIST-QRS-04-3031.
We are indebted to our collaborators in NOAA, FWRI and GSFC.